Active Learning Strategies for Technology Assisted Sensitivity Review

Graham McDonald[1], Craig Macdonald[2], Iadh Ounis[2]

University of Glasgow, G12 8QQ, Glasgow, UK
[1]g.mcdonald.1@research.gla.ac.uk
[2]firstname.lastname@glasgow.ac.uk

Abstract. Government documents must be reviewed to identify and protect any *sensitive* information, such as personal information, before the documents can be released to the public. However, in the era of digital government documents, such as e-mail, traditional sensitivity review procedures are no longer practical, for example due to the volume of documents to be reviewed. Therefore, there is a need for new technology assisted review protocols to integrate automatic sensitivity classification into the sensitivity review process. Moreover, to effectively assist sensitivity review, such assistive technologies must incorporate reviewer feedback to enable sensitivity classifiers to quickly learn and adapt to the sensitivities within a collection, when the types of sensitivity are not known *a priori*. In this work, we present a thorough evaluation of active learning strategies for sensitivity review. Moreover, we present an active learning strategy that integrates reviewer feedback, from sensitive text annotations, to identify features of sensitivity that enable us to learn an effective sensitivity classifier (0.7 Balanced Accuracy) using significantly less reviewer effort, according to the sign test ($p < 0.01$). Moreover, this approach results in a 51% reduction in the number of documents required to be reviewed to achieve the same level of classification accuracy, compared to when the approach is deployed without annotation features.

1 Introduction

At least 95 countries implement Freedom of Information (FOI) laws legislating that governments documents should be *open* to the public[1]. However, many such documents contain *sensitive* information, such as confidential or personal information and, therefore, FOI laws provide *exemptions* to prevent the release of such information. Government documents must, therefore, be sensitivity reviewed to ensure that no exempt information is released.

Historically, sensitivity review has been an exhaustive manual review of all documents being considered for release. However, in the era of born-digital documents such as e-mail, this purely manual review is not feasible [1], for example due to the volume of digital documents that are to be reviewed. Recently, automatic sensitivity classification algorithms have been shown to have potential for effectively identifying sensitive information in documents [2–5]. However, the potential consequences from the inadvertent release of sensitive information can be severe, for example if the identity of an informant is made public it can put the informant and their family at risk. Therefore, until automatic sensitivity classification is trusted, all documents that are to be released

[1] http://www.right2info.org/access-to-information-laws/access-to-information-laws

will continue to be manually reviewed. With this in mind, there is a need for appropriate protocols to integrate sensitivity classifiers into the review process to assist reviewers.

Technology assisted review (TAR), most notably associated with e-discovery [6, 7], is a process whereby human reviewers and an Information Retrieval (IR) system actively work together to identify relevant documents. The TAR protocol typically consists of two components, a key-word search system and a learning algorithm. Given a collection of documents and a *request for production*, e.g. "find all documents relating to ..", the TAR system formulates a query[2] to retrieve an initial pool of documents to be manually reviewed and labeled, or *coded*. The labeled pool is then used as a *seed set* to train the learning algorithm. The TAR protocol is then an iterative process where by the learner predicts the *k* most relevant unlabeled documents which the reviewer labels. The newly labeled documents are added to the training data and the algorithm is re-trained.

The TAR protocol can potentially be adapted to meet the needs of digital sensitivity review. However, in sensitivity review there is no equivalent to the request for production, since the types of sensitivity within the collection are not known *a priori*. Moreover, a judgment of sensitivity is often dependent on the context in which the information is produced and the time at which it is reviewed. Therefore, with this in mind, we propose to derive a representation of the sensitivities within a collection by having a reviewer annotate the specific text in a document that led to the reviewer's decision that the document is sensitive. Moreover, we propose to incorporate this reviewer feedback into the classification model to more quickly learn and adapt to the sensitivities within a collection at the time of review, while using minimal reviewing effort.

One possible strategy for integrating reviewer feedback into classification is active learning [8]. In active learning, the learning algorithm selects the order that documents are presented to a reviewer, with the aim of minimising the reviewer effort that is required to learn an effective classifier. Active learning has previously been shown to be an effective strategy for e-discovery TAR [6] and for *topic-oriented* text classification [9]. However, sensitivity is not topic-oriented [3] and, therefore, it is not obvious which active learning strategy is most appropriate for sensitivity classification.

In this work, we simulate the technology assisted sensitivity review process to present a thorough evaluation of active learning strategies for identifying sensitivities within a collection. We test two well-known *uncertainty sampling* active learning strategies from the literature and evaluate, as an active learning strategy, a *semi-automated text classification* [10] approach, that has previously been shown to be effective for increasing the cost-effectiveness of sensitivity reviewers [3]. Moreover, we show that by extending these approaches to incorporate reviewer feedback from sensitive text annotations, we can improve upon the *raw* active learning strategies to develop effective sensitivity classifiers more quickly, i.e. using less reviewer effort.

The contributions of this paper are two fold. Firstly, we provide the first thorough evaluation of active learning strategies for automatic sensitivity classification. Secondly, we present an active learning strategy that integrates reviewer feedback, from sensitive text annotations, to identify features of sensitivity that enable us to learn an effective sensitivity classifier (0.7 Balanced Accuracy) using significantly less reviewing effort, according to the sign test ($p < 0.01$). This approach resulted in a 51% reduction in the

[2] In active learning parlance, "query" usually refers to membership queries i.e. the system poses queries in the form of instances to be reviewed. In this work we use query in the IR sense, i.e. a textual passage used to retrieve relevant documents from an IR system. For membership queries we say that the system suggests documents to be reviewed.

number of documents that had to be reviewed to achieve the same level of classification accuracy, compared to when the approach was deployed without annotation features.

The remainder of this paper is structured as follows. Firstly, we present related work in Section 2, before presenting the active learning strategies that we evaluate in Section 3. We present our experimental setup in Section 4 and results in Section 5, before, finally, presenting our conclusions in Section 6.

2 Related Work

In this section we, firstly, present work relating to automatic sensitivity classification, before discussing technology assisted review and active learning later in the section.

The task of automatically classifying sensitive information that is exempt from release under Freedom of Information (FOI) laws was first introduced by McDonald *et al* [2]. In that work, the authors presented a proof-of-concept sensitivity classifier for identifying two FOI exemptions. In [2], the authors showed that text classification [11] can provide an effective baseline approach for sensitivity classification, achieving markedly above random effectiveness (0.7372 Balanced Accuracy). In [2], the authors also extended text classification with additional hand-crafted features, such as named entities of interest (e.g. politicians) and a subjective sentences count, which resulted in improved effectiveness for most of the reported metrics (e.g. + 5% F_2).

Feature engineering for sensitivity classification was subsequently investigated further by McDonald *et al.* [5]. In that work, the authors constructed document representations using word embeddings to capture semantic relations in the documents, such as *who said what about whom*. In [5], the authors evaluated the effectiveness of these semantic features compared with textual and syntactic features and found that combining semantic and textual features resulted in the largest increases in effectiveness, identifying ~10% more sensitive documents than the baseline approach.

Other works on sensitivity classification have, for example, investigated identifying sequences of sensitive text within documents [4] and selecting an appropriate classifier kernel for sensitivity [12]. However, the approaches mentioned thus far [2, 4, 5, 12] have evaluated sensitivity classification as a 1-shot batch supervised learning process, and therefore relied on there being a pre-judged representative collection with reliably labeled examples of the sensitivities within the collection. This can be problematic for sensitivity classification since, as previously mentioned in Section 1, the types of sensitivity in the collection are not known a priori. Therefore, differently from [2, 4, 5, 12], in this work we investigate how to incorporate reviewer feedback into the learning process to quickly learn an effective sensitivity classifier using minimal reviewing effort.

Berardi *et al.* [3] was the first work to investigate optimising the cost-effectiveness of sensitivity reviewers. In that work, the authors evaluated the effectiveness of a *utility-theoretic* [10] semi-automated text classification (SATC) approach, for sensitivity classification. The approach of Berardi *et al.* [10] addresses a scenario in which the underlying state-of-the-art classifier is not effective enough to meet a strict level of accuracy required within an organisation, e.g. reviewing for sensitivity within governments. The approach ranks documents by the expected gain in accuracy that a classification system could expect to achieve by having a reviewer correct mis-classified instances. Berardi *et al.* [3] found that their approach achieved substantial improvements in overall classification (+3% – +14% F_2). However, the authors concluded that these improvements

were much smaller than their approach had achieved for *topic-oriented* classification tasks, for example in [10]. In this work, we evaluate the utility-theoretic approach of Berardi *et al.* [3, 10]. However, differently from those works, which assume that the underlying classifier is state-of-the-art, we evaluate the approach as an active learning strategy to incorporate reviewer feedback into the underlying sensitivity classifier.

Moving on to technology assisted review (TAR), as previously stated in Section 1, TAR is an iterative process, whereby a learning algorithm selects batches of documents to be presented to a reviewer to be labeled. The labeled documents are then added to the current training data and the learner is re-trained. This iterative process continues until it is judged that sufficiently many relevant documents have been identified [6]. TAR has been applied to fields such as systematic review for evidence-based medicine [13], test collection construction [14] and, most notably, e-discovery [7, 15], where TAR has been shown to be more effective and more efficient than exhaustive manual review [16].

We believe that the TAR protocol can be adapted to meet the needs of sensitivity review. However, there are two noticeable differences in the objectives of TAR, for example in e-discovery, and reviewing for sensitivity. Firstly, the goal of TAR for e-discovery is to identify *close to* all the relevant documents in a collection while minimising the required reviewing effort [6], while in sensitivity review we must identify *all* sensitivities in any documents that are to be released to the public. Secondly, as previously stated in Section 1, there is no request for production, or *query*, in sensitivity review. Therefore, in this work we simulate TAR for sensitivity review to incorporate reviewer feedback into the TAR protocol to quickly learn to identity sensitivities from the reviewer feedback. Moreover, we evaluate approaches for selecting documents to be presented to a reviewer so that we can learn the sensitivities using the least reviewing effort possible.

Many TAR approaches deploy an active learning component to select documents to be reviewed. For example, Cormack and Grossman [6] presented an approach called *continuous active learning* and showed that selecting initial training documents through a simple keyword search, and subsequent training documents by continuous active learning, required significantly less (according to a sign test with $p < 0.01$) reviewing effort to achieve any given level of recall, compared to when the learning algorithm did not implement an active learning strategy to select the documents to use for training.

Pool-based active learning is a well known paradigm where by the learner selects documents to be reviewed, and labeled, from a pool of unlabeled documents. The most popular approach to pool-based active learning is uncertainty sampling, which has been extensively studied for developing text classification algorithms [8]. For example, Lewis and Gale [17] evaluated the effectiveness of uncertainty sampling, compared with relevance sampling and random sampling. In that work, the authors found that, for the same amount of labeling effort, uncertainty sampling usually resulted in the most effective classifier compared to the other approaches, when relevant documents are relatively abundant in the collection.

However, the selection of an appropriate active learning strategy is dependent on the nature of both the type of classification task and the task's objective [8]. Moreover, most of the research into active learning for text classification addresses a scenario in which there is a large collection of representative unlabeled examples available. Differently from that scenario, in this work, we investigate how quickly different active learning strategies can effectively learn a classifier for sensitivity classification when the types of sensitivities in a collection are not known a priori.

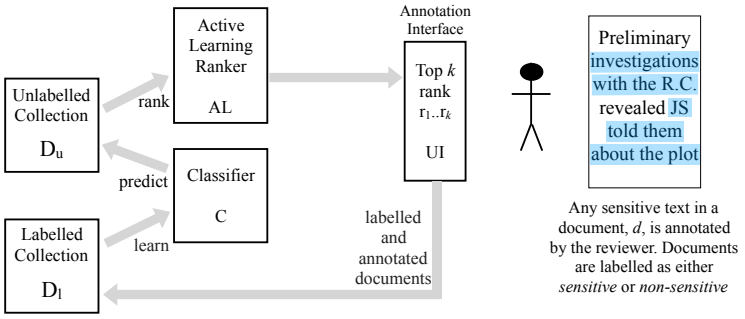

Fig. 1. Technology assisted sensitivity review simulation architecture.

3 Active Learning Methodologies

In this section, we present the active learning strategies that we evaluate for technology assisted sensitivity review. Firstly, in Section 3.1, we provide some preliminary information regarding our methodology for simulating technology assisted sensitivity review and the underlying classifier that we use as a basis for evaluating our active learning approaches. In Sections 3.2 - 3.4, we present the active learning strategies that we evaluate.

3.1 Preliminaries: Simulating Technology Assisted Sensitivity Review

Figure 1 presents the process that we deploy to simulate technology assisted sensitivity review. The process aims to efficiently solicit sensitivity judgments, for a document collection, D, that can subsequently be used to train a sensitivity classifier. The collection, D, consists of two separate subsets. Firstly, an unlabeled collection, D_u, for which we do not know the collection's sensitivities and, secondly, a labeled collection, D_l, which has been sensitivity reviewed and, therefore, has associated class labels, $l_i, l \in \{sensitive, nonSensitive\}$. Initially, $|D_u| = |D|, |D_l| = 0$ and, moreover, at all times $|D_u| + |D_l| = |D|$. Our review simulation consists of three separate system components. Firstly, an active learning component, AL. At each iteration of the review cycle, AL, is responsible for identifying k documents from D_u that would be likely to provide the most valuable evidence for training a sensitivity classifier, if their associated class labels, $l_1..l_k$, were known. To do this, AL ranks documents, $d_1..d_{|D_u|}, d_i \in D_u$, by means of an active learning strategy, al_j, and selects the top k ranked documents. These top k documents are presented to the reviewer in rank order, $d_1..d_k$, via the second system component, a user interface, UI, that enables the reviewer to label each of the documents with a corresponding class label l_i. For documents that are labeled $l_{sensitive}$, the reviewer also provides text-level annotations, $a_{di}, |a_d| \in \{0..|d_i|\}$, as illustrated in Figure 1, that indicate which text within the document led to the reviewer's l_i decision. The newly labeled documents, with their associated labels $l_1..l_k$ and annotations a_d are integrated into the labeled document set, D_l. Documents from D_l are then used to train the final system component, a sensitivity classifier, C. For C we select a multinomial naive Bayes (MNB) classifier, since it has been shown to be effective for text classification tasks [18] and, moreover, the model can be easily adapted to integrate different sources of feature evidence by simply weighting the underlying feature's multinomial [9, 19]. Once C has been trained, it is deployed and its predicted class labels, $\hat{l}_i$, with a corresponding confidence score, c_i, for the documents in D_u are input to

AL to provide evidence of the classifier's current knowledge. The simulation proceeds in this iterative cycle until all documents are labeled, $|D_u| = 0, |D_l| = |D|$.

3.2 Uncertainty Sampling

Uncertainty sampling [17] is a well known set of active learning approaches for evaluating the informativeness of documents in an unlabeled collection. In uncertainty sampling the algorithm tries to identify, and present to a reviewer, the documents in the collection for which the classifier is least certain about their correct class labeling.

In general, uncertainty sampling is a popular set of approaches for active learning since they are relatively easy to implement, are not computationally expensive and have been shown to be effective for many classification tasks [8]. Moreover, when deployed with a classifier that outputs probabilities or confidence scores, the classifier can be viewed as a *black box*. We test two well-known uncertainty sampling approaches, which have previously been shown to be effective for topic-based text classification [8, 17]. However, as previously mentioned in Section 1, sensitivity is not topic-based and, therefore, we can not presume that they will be effective for sensitivity classification.

The first uncertainty sampling strategy that we evaluate is *entropy based* uncertainty [8]. Entropy uncertainty sampling ranks documents by the sum of their label entropies [20], $H(L) = -\sum_i P(l_i) log P(l_i)$, over all possible labels, l_i. One way to view the intuition of this approach is that it calculates the number of bits it would take to encode the distribution of possible outcomes for L. Therefore, documents with a high $H(L)$ score should provide more information about their assigned label.

The second uncertainty sampling strategy that we evaluate is *margin* sampling [21], $M(d_i, l_1, l_2) = |P(l_1|d_i) - P(l_2|d_i)|$. This approach to uncertainty sampling calculates the margin, or difference, between the classifier predicted probability scores for a document's first and second most likely classification labels. The intuition of margin sampling is that documents with a small margin between the two most likely class prediction probabilities are more ambiguous and, therefore, knowing the class label of these documents would be most beneficial to the classifier.

3.3 Utility

As previously mentioned in Section 2, we evaluate the approach of Berardi *et al.* [10] as an active learning strategy for technology assisted sensitivity review. Berardi *et al.*'s approach was designed to rank documents in an order that would achieve the maximum increase in overall classification if a reviewer was to start from the top of the ranking and proceed down the list correcting any mis-classifications until an available reviewing budget had expired. This scenario is different from active learning in that it assumes that the underlying classifier is state-of-the-art and its objective is to produce the most effective ranking for a given reviewing budget. However, we believe that the utility-theoretic approach should perform well as an active leaning strategy for sensitivity classification since it has previously been shown to be able to improve the cost-effectiveness of sensitivity reviewers [3] and, moreover, by feeding the corrected classifications back into the learning process we are, in effect, just closing the loop in the active learning cycle.

The approach's intuition is that in text classification problems where there is an imbalance in the distributions of classification categories, and a metric is chosen to account for this imbalance (e.g. F_2), the improvements in effectiveness, or *gain*, that are derived

from correcting a false positive prediction is not the same as that for correcting a false negative prediction. This is important for sensitivity, since the consequences of misclassifying a sensitive document are much greater than that of a non-sensitive document.

In the case of binary classification, the utility-theoretic measure is defined as $U(d_i) = \sum_e P(e)G(e)$, where $P(e)$ is the probability of an event, e, occurring (i.e. a false negative or a false positive prediction) and $G(e)$ is the gain that can be obtained if that event does occur. To calculate the probability of an event occurring, the approach relies on the underlying classifier's label predictions, $\hat{l}_i$, on documents in D_u to be reliable. The probability of a false negative prediction, given that the classifier has made a negative prediction, is then calculated as $P(FN(d_i)|\hat{l}_i = neg) = 1 - \frac{e^{\sigma c_i}}{e^{\sigma c_i}+1}$, where $\frac{e^{\sigma c_i}}{e^{\sigma c_i}+1}$ is a generalised logistic function that monotonically converts a classifier's confidence score, c, in the range $(-\infty, +\infty)$ to real values in the range $[0.0, 1.0]$. The probability of a false positive occurring is computed analogously.

$G(e)$ is calculated on D_l and $G(FN) \neq G(FP)$. This inequality is reflected in the definitions of the gain functions $G(FN) = \frac{1}{FN}(\frac{2(TP+FN)}{2(TP+FN)+FP} - \frac{2TP}{2TP+FP+FN})$ and $G(FP) = \frac{1}{FP}(\frac{2TP}{2TP+FN} - \frac{2TP}{2TP+FP+FN})$. To compute $G(FN)$ and $G(FP)$ the TP, FP and FN frequency counts are derived by performing a *k*-fold cross validation on D_l. The corresponding frequencies are then obtained by the maximum-likelihood estimation $\hat{\alpha}^{ML} = \alpha^{D_l} \cdot |D_l|/|D_u|$, $\alpha \in \{TP, FP, FN\}$. Berardi *et al.* provide a thorough examination of the approach in [10], however it is worth noting that when calculating the $\hat{\alpha}^{ML}$ values, to avoid zero counts, Laplace smoothing is applied to each $\hat{\alpha}^{ML}$ in an *on-demand* fashion if any $\hat{\alpha}^{ML} < 1$, resulting in $\hat{\alpha}^{ML} + 1$.

3.4 Sensitivity Annotation Features

The active learning strategies presented in Sections 3.2 and 3.3 use predictions from the classifier, C, as evidence of the classifier's confidence in correctly classify the unlabeled documents, D_u. However, sensitive information is often only a small passage of text within a document and, therefore, we expect an active learning strategy that integrates term-level features of sensitivity to produce a more confident classifier that, in turn, will enable the active learning strategy to select more informative documents.

With this in mind, in this section, we present three strategies, inspired by Settles [9], that integrate term-level sensitivity features into the active learning process. As shown in Figure 1, when a document, d_i, is judged to be sensitive, the reviewer annotates the sensitive text within the document, $a_{di}, |a_d| \in \{0..|d_i|\}$. The strategies presented here utilise these document annotations to extend the strategies presented in Sections 3.2 and 3.3 with informative term-level sensitivity features.

We refer to our first annotation features strategy as *simple* annotation features. The simple strategy assumes that all the terms that a reviewer annotates are equally useful for identifying sensitivity. To integrate term feature importance into the active learning process, we simply increase the prior for the corresponding multinomial in the classifier, C, by a constant value α. This strategy is denoted as +Anno in Section 5.

The remaining two strategies make use of the labeled collection of documents, D_l, and the classifier's predictions on the unlabeled documents in D_u to calculate the expected information gain, $IG(f_k) = \sum_{F_k} \sum_i P(F_k, y_i) log \frac{P(F_k, y_i)}{P(F_k)P(y_i)}$, of term features in the unlabeled collection D_u, where $F_k \in \{0, 1\}$ indicates the presence or absence of a feature f_k in the class $y_i, y_i = l_i \cup \hat{l}_i$. The first information gain annotation features

strategy that we present considers all the term features that are in the intersection of the terms identified by $IG(f_k)$ and the terms annotated by a reviewer, in the current batch of documents being reviewed, as good sensitivity features and increases the prior for the corresponding multinomial in the classifier, C, by α. We refer to this strategy as *information gain* annotation features, denoted as +Anno$_{IG}$ in Section 5.

The final annotation features strategy that we evaluate, *annotation pool*, identifies useful sensitivity features through the same process as the previous information gain strategy, except that instead of only considering annotation terms from the current batch of documents being reviewed, a pool of potential sensitivity features is built from all previous annotations and any terms that are in the intersection of the terms identified by $IG(f_k)$ and terms in the annotation pool are considered as being good sensitivity features. This approach is denoted +Anno$_{POOL}$ in Section 5.

4 Experimental Setup

In this section, we present our experimental setup for evaluating the effectiveness of active learning strategies for technology assisted sensitivity review. We aim to answer two research questions, namely: **RQ1**; "Which active learning strategy enables the system to learn an effective sensitivity classifier with least reviewer effort?", and **RQ2**; "Which method of integrating a reviewer's annotations feedback is most effective for extending the tested active learning approaches?".

We evaluate our research questions on a test collection, T, of 3801 government documents that have been sensitivity reviewed by government sensitivity reviewers. The collection was assessed for two UK FOI exemptions, namely international relations and personal information. Any documents that contain any exempt information are labeled *sensitive*. The remaining documents are labeled *non-sensitive*, resulting in 502 sensitive documents (~13%) and 3299 non-sensitive (~87%).

To ensure the generalisability of our findings, we run our experiments over 25 stratified samples of the collection T. For each sample, we select 2500 documents from T as a training set Tr, which we use for the active learning simulation i.e. $|D_u| + |D_l| = Tr = 2500$. We select 500 documents from T as a held out test set, Te, for evaluating the performance of the classifier, C. We retain the distributions of sensitive and non-sensitive documents from T when generating Tr and Te, resulting in $Tr = \{2150$ non-sensitive, 325 sensitive$\}$ and $Te = \{435$ non-sensitive, 65 sensitive$\}$. We perform a binary classification, *sensitive* vs. *non-sensitive* and report mean scores over 25 samples. To test for statistical significance when evaluating reviewer effort, following [6], we use a sign test with $p < 0.01$.

At each iteration of the active learning cycle, we present the reviewer a new batch of *k* documents. For our experiments, we set $k = 20$. Previous work has shown that balancing the class distributions when training sensitivity classifiers can lead to a markedly improved model [2, 3, 5]. Therefore, when integrating newly labeled documents to D_l, we introduce the following constraint: |non-sensitive| $\in D_l \leq (k/2) +$ |sensitive| $\in D_l$. We discard documents that violate this constraint[3].

[3] In practice this means that we randomly down-sample the classifier's training data to loosely match the class frequencies. In preliminary experiments this led to uniform improvements across all tested approaches of ~+0.4 Balanced Accuracy, after all documents had been reviewed.

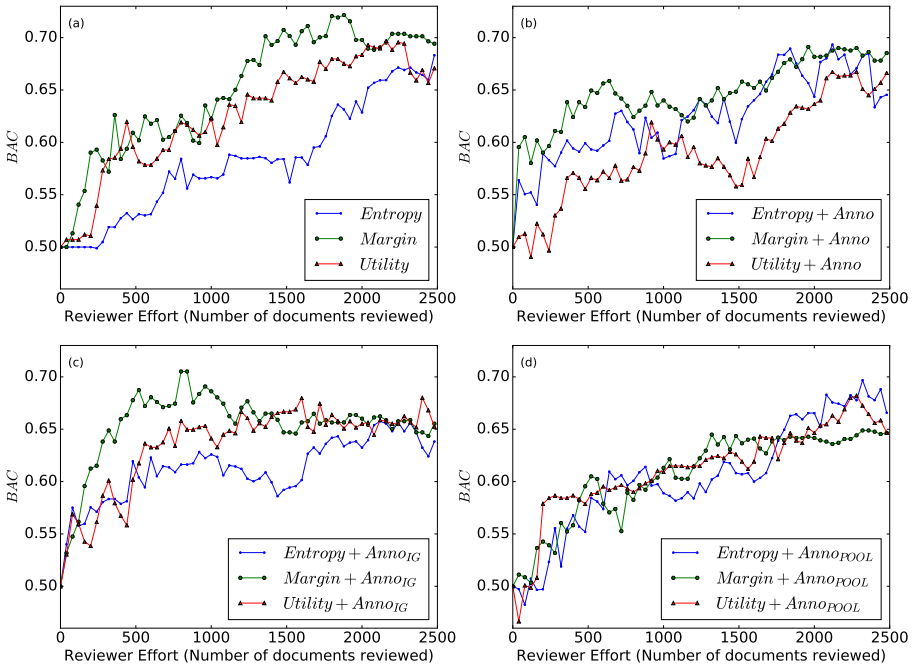

Fig. 2. Reviewer effort vs. Classifier effectiveness measured by Balanced Accuracy (BAC). Raw approaches are presented in (a), while (b) presents the approaches extended with *simple* annotation features, (c) presents the approaches extended with *information gain* annotation features, and (d) presents the approaches extended with *annotation pool* features.

For the utility approach, presented in Section 3.3, when estimating $G(FN)$ and $G(FP)$, following Berardi *et al.* [10], we select F_2 as our metric and perform a *k*-fold cross validation, setting $k = 10$. For the feature labeling approach, presented in Section 3.4, when integrating feature importance to the classifier, following [9], we set $\alpha = 50$.

5 Results

In this section, to answer the research questions presented in Section 4, we present the results of our active learning classification experiments. Figure 2 presents four plots that show the performance improvements of the learned classifier in terms of Balanced Accuracy (BAC), as evaluated on the held out collection Te. In each of the plots, the *x* axis shows the required reviewer effort, in number of documents reviewed. In Figure 2, plot (a) presents the results for the raw Entropy, Margin and Utility approaches, while plot (b) shows each of the approaches extended with the *simple* reviewer annotation features, plot (c) presents the approaches extended with *information gain* annotation features and, finally, plot (d) presents the approaches extended with *annotation pool* features.

Firstly, addressing **RQ1**, we evaluate the effectiveness of each active learning strategy for quickly learning a classifier that can reliably predict sensitivity. From Figure 2(a), we see that the Margin and Utility approaches begin to identify sensitivity noticeably quicker than Entropy, with Margin and Utility resulting BAC scores of 0.59 and 0.57 respectively when only 250 documents have been reviewed, while Entropy results

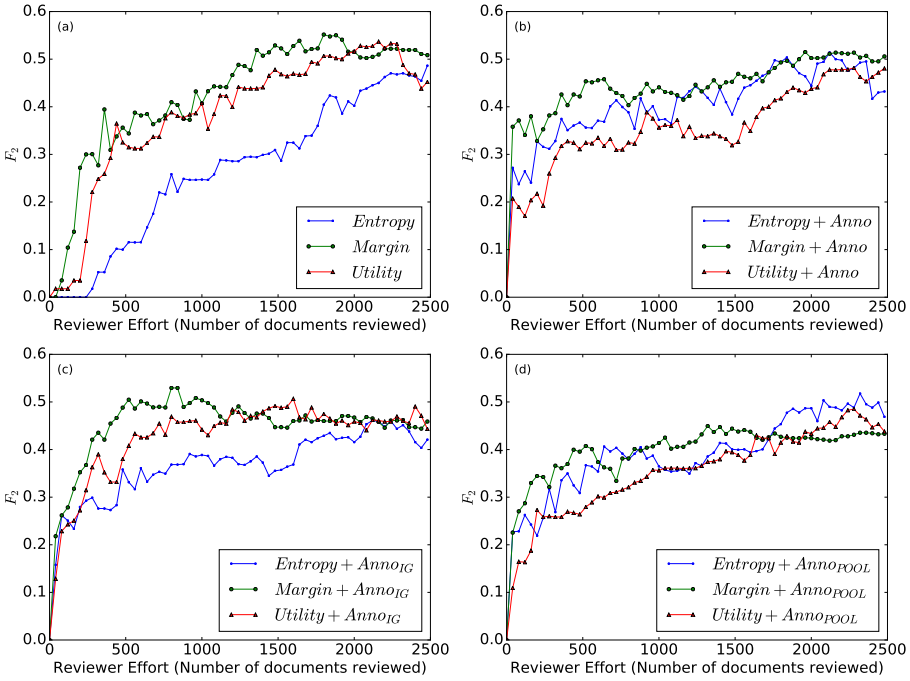

Fig. 3. Reviewer effort vs. Classifier effectiveness measured by F_2. Raw approaches are presented in (a), while (b) presents the approaches extended with *simple* annotation features, (c) presents the approaches extended with *information gain* annotation features, and (d) presents the approaches extended with *annotation pool* features.

in a random classifier (0.5 BAC). Moreover, the Margin and Utility approaches sustain this additional performance over Entropy for almost the entire review session. As the number of labeled documents increases, particularly when the number of reviewed documents is > 1180, we see that Margin shows noticeable improvements compared to the Utility approach. However, the difference between the approaches reduces as the number of reviewed documents approaches 2500. Therefore, in response to **RQ1**, we conclude that the Margin active learning strategy is the best performing strategy when the approaches are not extended with annotation features.

Turning our attention to **RQ2**, Figure 2(b),(c) and (d), present the active learning approaches with additional annotation features. From Figure 2(b), we see that the Entropy and Margin approaches with additional *simple* annotation features ($+Anno$) begin to identify sensitivity with markedly less reviewer effort than the approaches on their own (Figure 2(a)). To achieve 0.6 BAC, Margin + Anno required 200 documents to be reviewed while Margin required 400. Moreover, Entropy + Anno achieves 0.6 BAC with significantly less reviewing effort than Entropy, according to a sign test with $p < 0.01$ (400 documents vs. 1800 documents).

In evaluating the overall performance increase that is obtained from additional reviewer annotation features, we note from Figure 2(c) that *information gain* annotation features enable each of the approaches to develop an effective sensitivity classifier noticeably quicker than the raw approaches in Figure 2(a). Most notably, Margin sustains its initial gains in classification effectiveness and reaches its peak classification per-

Table 1. Area Under the Curve for the BAC and F_2 plots presented in Figures 2 and 3 respectively.

	BAC	F_2		BAC	F_2		BAC	F_2		BAC	F_2
Entropy	0.5800	0.2675	+ Anno	0.6213	0.4070	+ Anno$_{IG}$	0.6087	0.3674	+ Anno$_{POOL}$	0.6029	0.3924
Margin	0.6480	0.4271	+ Anno	0.6432	0.4454	+ Anno$_{IG}$	**0.6501**	**0.4503**	+ Anno$_{POOL}$	0.6022	0.3963
Utility	0.6236	0.3863	+ Anno	0.5871	0.3578	+ Anno$_{IG}$	0.6084	0.3551	+ Anno$_{POOL}$	0.6317	0.4262

formance (~0.7 BAC) with significantly less reviewer effort (according to the sign test, $p < 0.01$), requiring only 820 documents to be reviewed as opposed to 1700 when Margin is deployed without annotation features (shown in Figure 2(a)), therefore, resulting in a 51% reduction in required reviewer effort. However, we note that there is a notable decline in classification performance after this peak.

When classifying sensitive information, there is a much greater penalty from misclassifying documents that are sensitive than ones that are not. The F_2 metric reflects this asymmetry and, therefore, we present the classification improvements in terms of F_2 in Figure 3. We can see that the plots in Figure 3 display similar trends as the BAC plots, with Margin performing best and, moreover, information gain annotation features resulting in an effective classifier with notably less reviewing effort. Therefore, in response to **RQ2**, we conclude that information gain annotation features are most effective for integrating reviewer feedback from sensitivity annotations. We note, however, that the Utility approach is very competitive in terms of F_2 for the raw active learning approaches (Figure 3(a)) and when extended with information gain annotation features (Figure 3(c)). This is intuitive since the utility approach is optimised for F_2.

Finally, to provide a measure of overall classification effectiveness, Table 1 presents the Area Under the Curve (AUC) scores for the BAC and F_2 plots presented in Figures 2 and 3 respectively. As can be seen from Table 1, Margin + Anno$_{IG}$ achieves the best overall classification effectiveness throughout the review simulation. This finding provides extra evidence that the Margin + Anno$_{IG}$ combination can be an effective choice for technology assisted sensitivity review.

6 Conclusions

In this work, we presented a thorough evaluation of active learning strategies for technology assisted sensitivity review. We evaluated two well-known *uncertainty sampling* active learning strategies from the literature and an approach adapted from semi automated text classification, that has previously been shown to be effective for improving the cost-effectiveness of sensitivity reviewers. Moreover, we extended these approaches to integrate term-level reviewer feedback from annotations of sensitive text within documents. We showed that extending Margin uncertainty sampling with high information gain annotation term features enabled us to learn an effective sensitivity classifier (0.7 BAC) using significantly less reviewing effort (according to the sign test with $p < 0.01$), than when the approach was deployed without annotation features, i.e. a 51% reduction in the number of documents that had to be reviewed. Moreover, we found that this approach achieved the best overall classification effectiveness throughout a technology assisted sensitivity review simulation, and conclude that the approach can be an effective choice for quickly learning to classify sensitivity, when the types of sensitivities in a collection are not known *a priori*.

References

1. TNA: The application of technology-assisted review to born-digital records transfer, inquiries and beyond (2016)
2. McDonald, G., Macdonald, C., Ounis, I., Gollins, T.: Towards a classifier for digital sensitivity review. In: Proc. ECIR. (2014)
3. Berardi, G., Esuli, A., Macdonald, C., Ounis, I., Sebastiani, F.: Semi-automated text classification for sensitivity identification. In: Proc. CIKM. (2015)
4. McDonald, G., Macdonald, C., Ounis, I.: Using part-of-speech n-grams for sensitive-text classification. In: Proc. ICTIR. (2015)
5. McDonald, G., Macdonald, C., Ounis, I.: Enhancing sensitivity classification with semantic features using word embeddings. In: European Conference on Information Retrieval, Springer (2017) 450–463
6. Cormack, G.V., Grossman, M.R.: Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, ACM (2014) 153–162
7. Oard, D.W., Baron, J.R., Hedin, B., Lewis, D.D., Tomlinson, S.: Evaluation of information retrieval for e-discovery. Artificial Intelligence and Law **18**(4) (2010) 347–386
8. Settles, B.: Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning **6**(1) (2012) 1–114
9. Settles, B.: Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In: Proc. EMNLP. (2011)
10. Berardi, G., Esuli, A., Sebastiani, F.: A utility-theoretic ranking method for semi-automated text classification. In: Proc. SIGIR. (2012)
11. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1) (2002) 1–47
12. McDonald, G., García-Pedrajas, N., Macdonald, C., Ounis, I.: A study of svm kernel functions for sensitivity classification ensembles with pos sequences. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '17, New York, NY, USA, ACM (2017) 1097–1100
13. Lefebvre, C., Manheimer, E., Glanville, J.: Searching for studies. Cochrane handbook for systematic reviews of interventions: Cochrane book series (2008) 95–150
14. Sanderson, M., Joho, H.: Forming test collections with no system pooling. In: Proc. SIGIR. (2004)
15. Oard, D.W., Hedin, B., Tomlinson, S., Baron, J.R.: Legal track overview. In: Proc. TREC. (2008)
16. Grossman, M.R., Cormack, G.V.: Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. Rich. JL & Tech. **17** (2010) 1
17. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Proc. SIGIR. (1994)
18. Rennie, J.D., Shih, L., Teevan, J., Karger, D.R.: Tackling the poor assumptions of naive bayes text classifiers. In: Proc. ICML. (2003)
19. McCallum, A., Nigam, K., et al.: Employing em and pool-based active learning for text classification. ICML **98** (1998) 350–358
20. Shannon, C.E.: A mathematical theory of communication. BSTJ **27** (1948) 623–656
21. Scheffer, T., Decomain, C., Wrobel, S.: Active hidden markov models for information extraction. In: International Symposium on Intelligent Data Analysis, Springer (2001) 309–318